SUPPLEMENTARY INFORMATION

Phase 1: Pre-processing details. In case of genomic sequences, repeats were masked with 'N's or excluded beforehand in order to avoid clusters made of genomic repeats. Contiguous strings with more than 15 'N's were deleted and the resulting fragments were treated as independent sequences. Long sequences were split into smaller fragments (reasonable fragment sizes are between 100-250 nt) to detect local signals. We performed the split only if both fragments were both longer than a required minimal length. We removed near identical sequences with BLASTCLUST (Altschul *et al.*, 1997). Such sequences would pollute the clusters and would overshadow more subtle sequence-structure relationships. We identified clusters of sequences which are more than 90% identical over 90% of the sequence length. From each such cluster we kept only one sequence at random and removed all the others. The removed sequences are however included in the final clustering results (see Phase 9). Filtering with BLASTCLUST was applied iteratively until no sequence duplicates were found.

`GraphClust` parameters. If not stated differently in the main paper, the following parameters were applied while running `GraphClust`. Phase 2: RNASHAPE abstraction level: 3. Phase 4: Benchmark sets and human lincRNAs were clustered without sampling. For RNAZ screens and the human EVOFOLD set a random sample of 50% was used. For human 3'UTR set and Fugu lincRNAs a sample of 20% was used. The initial hash signature used 300 hash functions. Iteratively, the signature size was increased by steps of 50 hash functions. No overlap between the returned dense regions was allowed. Phase 7: For benchmark sets hits with a bitscore $\geq$ 15 were considered as significant, for all the other datasets the threshold was $\geq$ 20.

`GraphClust` space and memory requirements. A memory limit of 3.5GB was set for all `GraphClust`-phases except phase 4. The memory requirement in phase 4 depend directly on the size of the sparse vector. For example, the sparse vector of the Rfam benchmark is ≈500MB, for the Fruit fly RNAz screen ≈1,7GB and for the human 3'UTR ≈15GB. Assuming sequence fragments of similar size, the sparse vector increases linear in the number of sequences. Datasets up to 30.000 sequence fragments (≈150nt) are therefore possible to process on a normal machine (with 4 GB RAM).

`GraphClust` software and hardware. The pipeline is implemented in Perl. Other used tools: LOCARNA (v1.6.2), Vienna RNA package (v1.8.5), RNAclust/ RNASOUP (v1.2.5), RNASHAPE (v2.1.6), INFERNAL (v1.0.2), BLASTCLUST (v2.2.15). The times were measured on Opteron 2356 (2.3 GHz) machines. For parallization the *Sun Grid Engine* (SGE) is used.

Table S1. `GraphClust` results for the Rfam benchmark set after each iteration for all predicted clusters for different partition types. The F measure is the average over all cluster predicted until the given iteration. The Rand index is based on the partition for all clustered RNA candidates. For the MERGED partition we provide in addition accuracy, precision and recall.

		Partition														
		BEST			MERGED						ORACLE			SOFT		
Iteration	#Seq	#C	F	Rand	#C	F	Rand	Accuracy	Precision	Recall	#C	F	Rand	#C	F	Rand
1	271	10	0.545	0.376	5	0.882	0.888	0.998	0.912	0.869	5	0.882	0.888	10	0.938	0.215
2	629	20	0.649	0.699	14	0.834	0.932	0.997	0.891	0.814	13	0.877	0.932	20	0.888	0.372
3	1076	30	0.743	0.864	23	0.868	0.956	0.997	0.928	0.841	22	0.894	0.957	30	0.879	0.429
4	1737	40	0.778	0.956	33	0.872	0.984	0.996	0.950	0.834	31	0.908	0.985	40	0.878	0.612
5	1844	50	0.774	0.943	42	0.860	0.984	0.997	0.924	0.839	39	0.906	0.985	50	0.862	0.609
6	2019	60	0.783	0.780	50	0.879	0.985	0.998	0.927	0.867	47	0.918	0.986	60	0.867	0.549
7	2181	69	0.786	0.783	58	0.877	0.985	0.998	0.922	0.875	55	0.910	0.986	69	0.865	0.554
8	2321	79	0.787	0.781	67	0.873	0.985	0.998	0.911	0.883	62	0.912	0.986	79	0.853	0.558
9	2391	89	0.782	0.782	77	0.856	0.985	0.998	0.892	0.872	70	0.908	0.986	89	0.838	0.559
10	2440	99	0.773	0.781	86	0.845	0.985	0.998	0.881	0.868	77	0.904	0.983	99	0.824	0.555
11	2503	109	0.767	0.779	94	0.846	0.984	0.998	0.878	0.872	84	0.907	0.983	109	0.821	0.555
12	2587	118	0.765	0.659	102	0.843	0.985	0.999	0.876	0.869	91	0.907	0.984	118	0.815	0.481
13	2671	128	0.768	0.661	112	0.839	0.985	0.999	0.868	0.869	100	0.902	0.984	128	0.807	0.482
14	2742	138	0.771	0.661	122	0.837	0.984	0.999	0.868	0.867	109	0.900	0.984	138	0.801	0.482
15	2821	148	0.766	0.649	130	0.834	0.984	0.999	0.856	0.876	117	0.892	0.983	148	0.796	0.483

Table S2. Aggregated serial time for Rfam benchmark set. Time_C denotes the average time per predicted candidate cluster C up to iteration i. Time_{ALL} is the total serial. All times are given in seconds.

i	#C	Phase 2	Phase 3		Time_i	Time_{ALL}	Time_C
0		4169	4145		8314	8314	
		Phase 4	Phase 5-6	Phase 7			
1	10	458	10633	3904	14995	23309	2331
2	20	416	17980	1564	19962	43272	2163
3	30	334	13239	1533	15108	58380	1946
4	40	260	11694	752	12708	71088	1777
5	50	186	11351	783	12321	83409	1668
6	60	171	8667	726	9566	92975	1549
7	70	154	11069	739	11964	104940	1499
8	80	133	3671	597	4402	109342	1366
9	90	120	3841	620	4581	113924	1265
10	100	114	2768	707	3590	117515	1175
11	110	108	2544	492	3145	120660	1096
12	120	99	1917	712	2728	123388	1028
13	130	90	1197	640	1929	125318	964
14	140	83	1220	512	1817	127135	908
15	150	77	1776	636	2491	**129626**	864

Table S3. `GraphClust` results for the small ncRNA benchmark set after each iteration for all predicted clusters for different partition types. The F measure is the average over all clusters predicted until the given iteration.

		Partition														
		BEST			MERGED						ORACLE			SOFT		
Iteration	#Seq	#C	F	Rand	#C	F	Rand	Accuracy	Precision	Recall	#C	F	Rand	#C	F	Rand
1	140	10	0.942	0.945	10	0.942	0.945	0.996	0.924	0.974	10	0.942	0.945	10	0.942	0.897
2	232	20	0.926	0.939	20	0.926	0.939	0.996	0.903	0.971	20	0.926	0.939	20	0.916	0.892
3	270	26	0.936	0.935	26	0.936	0.935	0.996	0.920	0.970	26	0.936	0.935	26	0.928	0.894
4	298	30	0.925	0.922	30	0.925	0.922	0.996	0.902	0.971	29	0.929	0.906	30	0.907	0.874
5	305	32	0.909	0.918	32	0.909	0.918	0.996	0.880	0.973	30	0.925	0.900	32	0.892	0.872
6	319	34	0.897	0.904	34	0.897	0.904	0.996	0.861	0.974	32	0.911	0.887	34	0.881	0.862
7	329	35	0.890	0.897	35	0.890	0.897	0.995	0.851	0.975	33	0.903	0.881	35	0.875	0.857
8	332	36	0.883	0.898	36	0.883	0.898	0.995	0.844	0.966	34	0.895	0.882	36	0.866	0.856
9	335	37	0.882	0.901	37	0.882	0.901	0.995	0.840	0.967	35	0.894	0.884	37	0.863	0.854
10	339	38	0.867	0.891	38	0.867	0.891	0.995	0.831	0.948	36	0.878	0.875	38	0.852	0.846
11	345	39	0.871	0.899	39	0.871	0.899	0.995	0.839	0.946	37	0.881	0.888	39	0.848	0.828
12	349	40	0.868	0.900	39	0.873	0.894	0.995	0.848	0.940	38	0.876	0.889	40	0.839	0.815
13	353	41	0.868	0.902	39	0.871	0.891	0.994	0.848	0.934	38	0.892	0.893	41	0.837	0.794
14	357	42	0.854	0.809	39	0.872	0.886	0.995	0.848	0.936	38	0.893	0.888	42	0.836	0.646
15	360	43	0.843	0.794	39	0.858	0.866	0.993	0.841	0.916	38	0.893	0.874	43	0.827	0.636

Table S4. Aggregated serial time for small ncRNA benchmark set. Time_C denotes the average time per predicted candidate cluster C up to iteration i. Time_{ALL} is the total serial. All times are given in seconds.

i	#C	Phase 2	Phase 3		Time_i	Time_{ALL}	Time_C
0		225	495		720	720	
		Phase 4	Phase 5-6	Phase 7			
1	10	41.76	2192	200	2434	3154	315
2	20	27.01	3132	236	3395	6549	327
3	26	17.02	7548	116	7681	14230	547
4	30	13.33	3618	163	3795	18025	601
5	32	8.63	3792	75	3876	21902	684
6	34	6.78	984	43	1034	22936	675
7	35	4.81	221	24	250	23186	663
8	36	3.65	230	21	255	23441	651
9	37	3.07	135	16	155	23596	638
10	38	2.74	102	24	129	23725	624
11	39	2.17	73	33	108	23832	611
12	40	1.73	62	23	87	23920	598
13	41	1.42	114	27	143	24063	587
14	42	1.15	115	30	146	24209	576
15	43	0.88	73	18	92	**24301**	565

Table S5. `GraphClust` results for the Rfam benchmark set and the small ncRNA benchmark using different sample sizes during phase 4. Shown is the result after 15 iterations. For the Rfam set the final quality is nearly identical between all used samples sizes. The small ncRNA set shows the same quality in terms of F measure and Rand index. Only the final number of clusters and clustered sequences is decreased. This is probably due to the small size of the dataset. For practical purposes this would mean to run 1-2 additional iterations of the pipline.

			Partition														
			BEST			MERGED						ORACLE			SOFT		
Sample	Iteration	#Seq	#C	F	Rand	#C	F	Rand	Accuracy	Precision	Recall	#C	F	Rand	#C	F	Rand
Rfam benchmark																	
100%	15	2821	148	0.766	0.649	130	0.834	0.984	0.999	0.856	0.876	117	0.892	0.983	148	0.796	0.483
50%	15	2707	150	0.758	0.655	127	0.850	0.984	0.999	0.888	0.881	117	0.909	0.988	150	0.809	0.479
25%	15	2754	149	0.758	0.636	130	0.835	0.990	0.998	0.867	0.877	120	0.888	0.995	149	0.799	0.518
small ncRNAs benchmark																	
100%	15	360	43	0.843	0.794	39	0.858	0.866	0.993	0.841	0.916	38	0.893	0.874	43	0.827	0.636
50%	15	299	38	0.854	0.812	35	0.880	0.913	0.994	0.913	0.890	33	0.920	0.925	39	0.856	0.698
25%	15	269	34	0.854	0.905	33	0.846	0.880	0.991	0.918	0.843	30	0.907	0.907	34	0.837	0.838

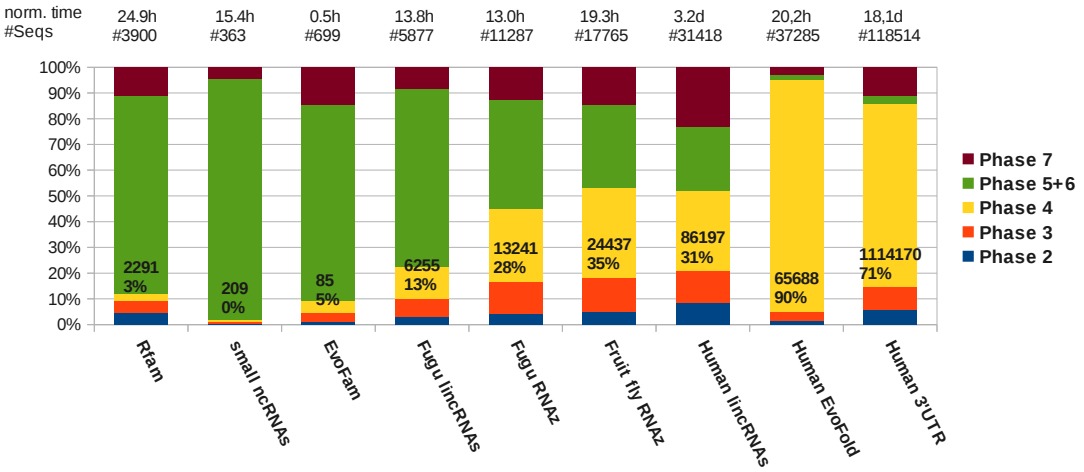

Fig. S1: **Runtime comparison for all analyzed datasets.** Shown is the relative runtime spent in phases 2-7 of the `GraphClust`-pipeline. Pre- and post processing phases are skipped. For the ease of comparison, we normalized all times to 5 iterations and 100 clusters. The time for phase 4 (clustering) is normalized to 100% sample size and is indicated on each bar (percentage and in seconds). On top of each dataset the normalized serial time and the number of sequences is given. For small datasets, the runtime is dominated by the cluster refinement step (Phase 5+6) which uses costly sequence-structure alignment. Please note that we do not normalize the influence of the sequence length which effects all phases, e.g. the RNA folding. The Rfam set and the small ncRNA set contain sequences up to 400 nt, whereas the EvoFam and EvoFold set contain mainly short sequences (average length = 36nt). RNAz screens have an average length of 120 nt and all other sets have an averge sequence length of 150 nt.

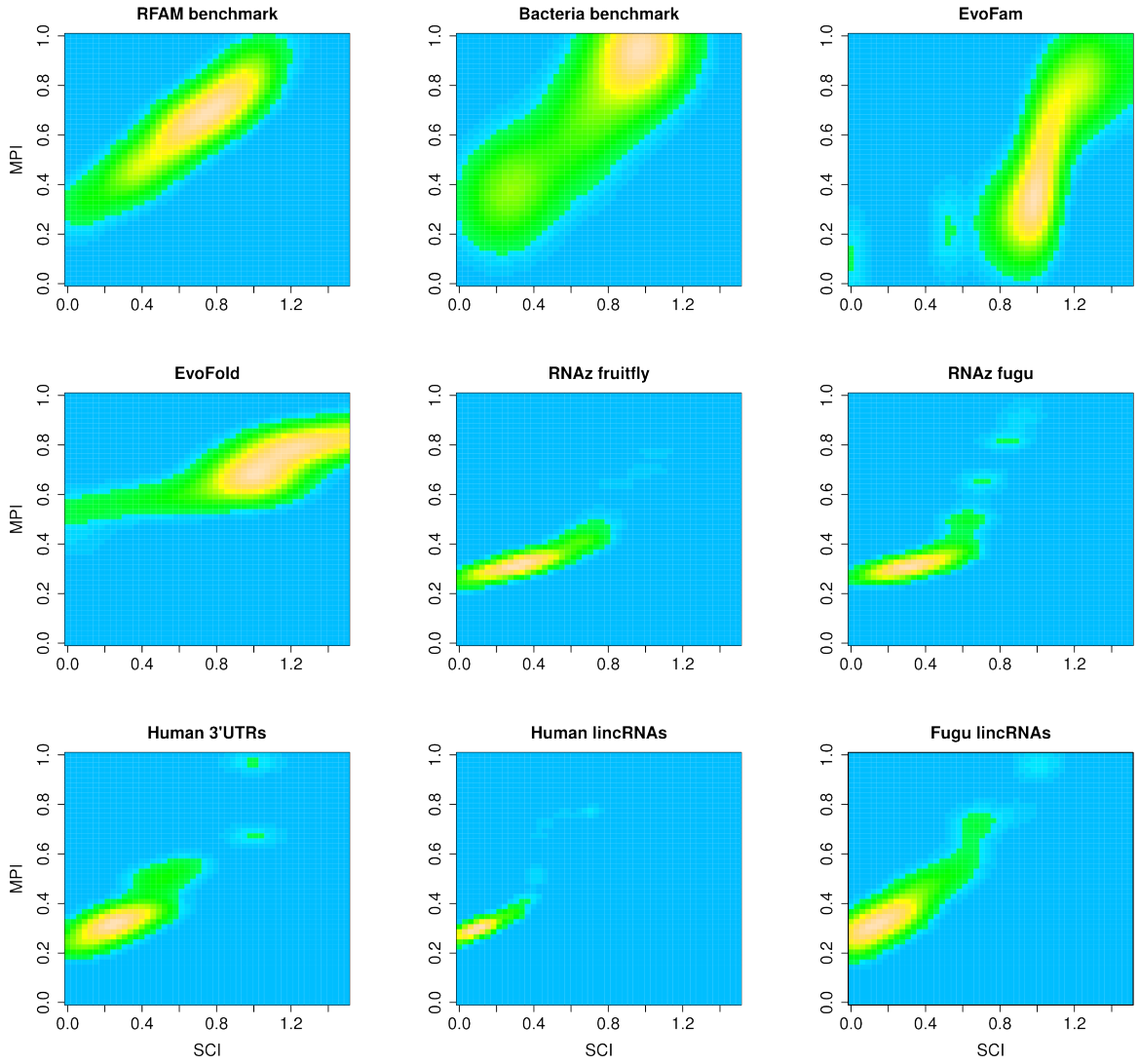

Fig. S2: **SCI/MPI density heat-maps of `GraphClust`-generated clusters.** The heat-maps illustrate that `GraphClust` can indeed identify local structural clusters. We present heat-maps for different benchmark and application scenarios. Recall that for local motifs the structure conservation index (SCI) can only be used as a measure of "structured-ness" in case it is high. Low SCIs are known to be uninformative for local structural elements and no conclusion can be drawn. Thus, although the mean pairwise sequence identity (MPI) is low for many structural clusters, we still observe clusters with reasonable high SCIs indicating conserved secondary structural elements.

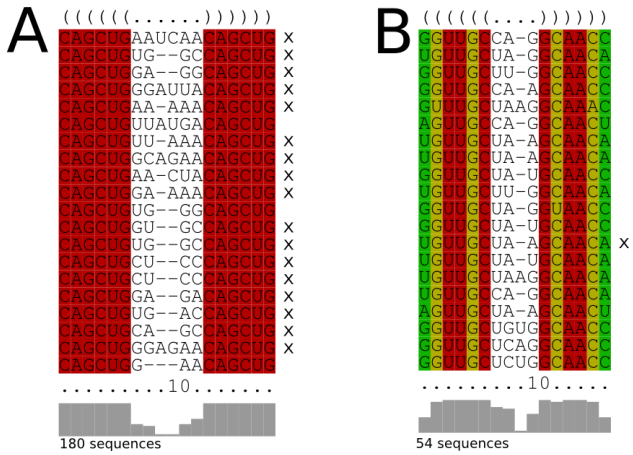

Fig. S3: **Two exemplary cluster identified by `GraphClust` when processing EvoFold hits.** For each cluster the top 20 sequences are given. The consensus secondary structures of both clusters are small hairpins. Cluster (A) contains many sequences that belong to the same EvoFam family (as indicated by 'x'). Contrary, only one of the depicted sequences of cluster (B) is a member of a previously described EvoFam family. Interestingly, this novel cluster contains several compensatory mutations that support the structural clustering. This demonstrates that `GraphClust` can identify relevant local structural clusters. It may not only help to improve existing family assignments, it can also be used to define new ones.

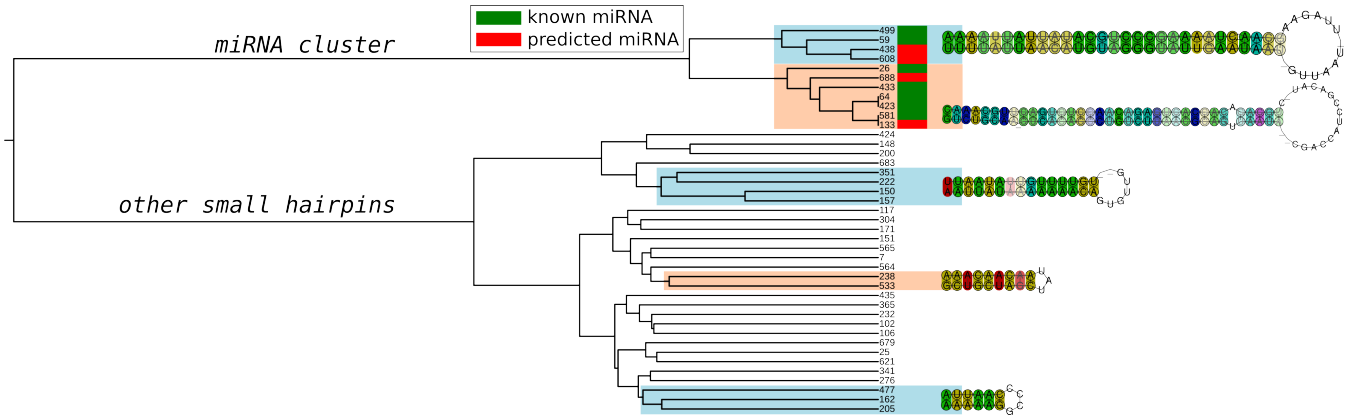

Fig. S4: **`GraphClust` identifies novel human miRNA candidates.** A hierarchical LocARNA-based structural clustering based on one representative sequence selected from `GraphClust`-derived clusters of EvoFam sequences reveals two main structural classes. Apart from several small hairpins, we observe a prominent miRNA cluster consisting of known miRNAs (annotated by miRBase v.17, highlighted in green) and structurally related sequences lacking any annotation (red). These are promising candidates for novel miRNAs.